Harmonization Sometimes Harms

Klenner, Manfred ; Göhring, Anne ; Amsler, Michael

Abstract: In this paper we argue that harmonization is not the preferred way to produce a gold standard in all cases. Neither does a majority vote based harmonization produce an appropriate gold standard centroid, nor would a mere centroid be a good basis for training a system that reproduces prototypical user reactions given some understanding task. We discuss these claims in the context of sentiment inference.

Harmonization Sometimes Harms

Manfred Klenner and Anne Göhring and Michael Amsler
Department of Computational Linguistics
University of Zurich, Switzerland
{klenner,goehring,mamsler}@cl.uzh.ch

Abstract

In this paper we argue that harmonization is not the preferred way to produce a gold standard in all cases. Neither does a majority vote based harmonization produce an appropriate gold standard centroid, nor would a mere centroid be a good basis for training a system that reproduces prototypical user reactions given some understanding task. We discuss these claims in the context of sentiment inference.

1 Introduction

It is common practice to harmonize annotated data produced by a couple of human raters in order to create a gold standard. The quality of the annotated data not only depends on the quality of the annotation guidelines, but also on various personal traits of the raters (cognitive capacity, reliability, motivation etc.). We cannot fully control these parameters, we are even expecting raters to fail and produce wrong annotations. Hence the need for harmonization where the right annotation decision is fixed either on the basis of a discussion among raters or by (simple) majority vote. This process of harmonization is suited for all those annotation tasks that have clear decision boundaries. The situation changes when the annotation is more based on subjective understandings and evaluations of the material (e.g. text). This might even touch upon personal standards, mental dispositions and ethic obligations not to mention political stance or religious premises. To give quite a harmless example: is the occasional (or even unique) use of a work phone for private purposes negative? Clearly, we could force raters to penalize any misuse, any (white) lies, and misdemeanor.

Would a gold standard created that way represent any real opinion or would it be an artifact of the guidelines.

In this paper, we argue that for our application, sentiment inference, we should leave room for individual perspectives and should avoid harmonization and accept that the annotation process leaves us with a distribution of annotations representing a diversity of opinions. We first introduce the notion of sentiment inference, then we sketch our annotation guidelines and discuss lessons learned from the initial annotations efforts.

2 Sentiment Inference

Sentiment inference[1] is a variant of stance detection where both the source and the target of an attitude need to be found and where the pro and con relations sometimes are only transitively given. In stance detection, the source of an opinion usually is the writer of a text and the target is some controversial topic (e.g. death penalty). However, any text might discuss sources and targets of attitudes, the text author just tells us about it and has or has not an opinion directed towards these proponents and opponents. For instance, a statement like *Obama criticizes the spread of fake news on Facebook* gives rise to the inference that the opinion source *Obama* is against (con) the target *fake news*. This is also an example of an inference, since literally, Obama just is against the spread, but this implies he is against fake news as well. The kind of reasoning that gives rise to *con(Obama, fake news)* could be captured by the following inference rule: If A is against B and B is good for C, then A is against C. Of course, one has

[1] We have implemented a rule-based baseline system for German, a demo is available under pub.cl.uzh.ch/demo/stancer/

to accept the idea that it is good for the fake news that it was spread. Then, we no longer only are talking about attitude, but also about positive and negative effects on entities. Effects are defined as the perceived consequences of an event that happened (this is similar to the good-for/bad-for distinction of (Deng et al., 2013)). Sometimes, there are only effects, but no attitudes (*she wins*) and sometimes attitudes and effects are even somewhat contrary (*He criticizes that she was honored*).

3 Sentiment Inference Annotation

Initially, our goal was to produce a traditional gold standard for verb-based sentiment inference. We have a lexicon of about 1000 verbs that express polar relations and for a subset of 100 verbs, we extracted 500 sentences from a newspaper corpus. We let 4 raters (experts) annotate them. The annotators were trained beforehand on 100 sentences from another newspaper corpus. On that basis, the initial annotation guidelines were refined as a result of our discussions of difficult examples and borderline cases.

vote(s)	# annot.	%	eff	rel	act
4	340	18.84	162	169	9
3	325	18.01	151	161	13
2	358	19.83	123	201	34
1	782	43.32	255	448	79

Table 1: 1805 different annotations in total 3833 annotations of 500 sentences

The distribution of annotations produced is shown in table 1. It also gives the absolute frequencies for effect (eff), relations (rel) and actors (act) (we won't discuss this last dimension here, because only a few cases were found).

In 43.32% of the cases only a single annotator (last row, 1 vote) produced a particular annotation in contrast to the other ones. In 19.83% we have 2 annotators that agree, in 18.01% three agree and in 18.84% of the cases all agreed. This is quite a diverse picture and we used Fleiss' Kappa to further quantify it. For the reached interannotator agreement see table 2.

The result was an agreement of 11.98% (overall value), which is considered slight (0.0-0.20 is slight, 0.21-0.40 is fair). We also measured the pairwise agreements which are -0.16%, 6.94%, 8.60%, 8.62%,14.85% and 27.02%. Only one pairing reached a fair agreement (27.02%).

	Fleiss' K	type	#
overall	11.98%	slight	1805
relation	10.36%	slight	979
- pro relation	7.38%	slight	383
- contra relation	12.27%	slight	596
effect	15.59%	slight	691
- positive effect	12.51%	slight	276
- negative effect	16.73%	slight	415
actor:	-5.05%	(poor)	135

Table 2: IAA for all 4 annotators

The agreement in general is low. Searching for the reasons, we again discussed our guidelines (see the next section), but in the end, after we finished our attempt to harmonize, we opted against more rigid guidelines and in favor of a more liberal notion of what is called a gold standard.

4 Sketch of the Annotation Guidelines

We distinguish two polar relations (pro, con), positive and negative effects, and positive as well as negative actors. According to our guidelines, a pro relation is directed from a source toward a target, if there is

1. a positive attitude of an actor toward an actor/object/situation
2. an action of an actor which yields something positive for an actor/object/situation
3. an object which yields something positive for an actor/object/situation

The con relation is defined accordingly.

We allow pro/con relations to hold between non-animate discourse referents, thus these relations are not meant to be interpreted as strict attitudes where the source must be an (intentional) opinion bearer. The reason is, among others, that in order to be able to infer pro/con relations transitively, non-animate arguments are useful as a bridge. In *The CEO is against the contract, since the contract is bad for the company* we can only infer that *pro(CEO,company)* if *con(CEO,contract)* and *con(contract,company)*. But *contract* is not a well-defined attitude bearer. This is why we have widened the definition of pro/con. If a given application needs a stricter perspective, an additional animacy classifier could be used to separate cases where the opinion source is animate - even metonomically given like in

Moscow criticizes Washington - from cases with non-animate sources like in *The snow blocks the entrance to the hospital*.

The guidelines also state that the pro and con relations are situation specific, they do not hold in general. If A criticizes B, then this is true only for the situation at hand.

Effects are consequences of the truth of a particular situation. They hold if the event denoted by a clause is factual (cf. a positive effect on *she* given *She won the competition*, but not in *She might win*). Like attitudes, effects are verb-specific. We distinguish on a conceptual level (but we do not annotate it) moral (*to accuse*), social (*to honor*), emotional (*to insult*) and physical (*to hurt*) effects.

5 Disagreement Example

In order to give an example of the problems we encountered, we discuss the following (translated version of a German) sentence:

Jim Crace, whose books depress many readers, seemed like the most cheerful man on earth.

All annotators agreed that there is a negative effect on 'readers' (1). Three see a con relation between 'book' and 'readers' (2), one believes in a con relation between 'readers' and 'Jim Crace' (3), one opts for a con between 'readers' and 'book' (4) and one annotator postulates a negative effect on 'Jim Crace' (5). In the harmonization process, the annotation 1 and 2 survived; the annotator of 3 sticks with her opinion; it was unproblematic to cancel 5; there were longer discussions on relation 4 that was finally given up (by its proponent).

Clearly as a reader one immediately has a strong opinion here, but there are so many aspects to consider if you start a discussion on such cases – it is really amazing.

6 Disagreement Analysis

For 50 sentences, we performed a disagreement analysis. This means, we harmonized our annotations to get a gold standard, thereby checking where and why our annotations diverged.

In total, the four annotators produced 432 individual annotation decisions for the 50 sentences, which amounts to 204 distinct annotations. After harmonization, we ended up with 113 gold labels. Sometimes all 4 annotators agreed upon a particular decision, then 4 of the 432 individual annotations yield a single gold standard annotation (one of 113). Counting on individual annotations, there was a 100% agreement (all four agreed) for 37 annotations (18% of all annotations). See table 3 for a detailed description. A comparison with the whole sample (see table 1) reveals that our 50 sentences sample (table 3) well reflects the underlying distribution of classes and thus might be considered as representative.

votes	**4**	**3**	**2**	**1**
annotations	37	39	39	89
%	18	19	19	44

Table 3: Distribution of annotations (50 sentences)

The main question was: could we harmonize the data without (much) dispute and discord? This clearly would be the case if individual mistakes are the reason for disagreement.

It turned out that just 6% of the annotations (27 out of 432) are based on plain mistakes (e.g. wrong head of a phrase) and thus could be resolved immediately. For another 34% of the annotations, the annotators agreed during the discussion that they missed out on an annotation (87 cases) or that their annotation was wrong (60 cases out of 432). So 40% of the disagreement was away and here there was hardly the need for strict argumentation to convince each other to accept an additional annotation or to drop one.

But there are also cases where no agreement was reached, i.e. there was at least one annotation in a sentence which not all four annotators could agree on. There are two cases: 26 out of 432 (6%) annotations on which not all four agreed and 6 cases (1.4%), where one annotator did not agree on an annotation which the others proposed.

The rest of the cases (53.4%) are cases where only after some discussion a harmonization step was carried out. While 40% of the cases are valid harmonizations, the nature of the harmonization of the remaining 53.4% cases is unclear. In the next two sections, we argue that strict harmonization is harming (for our task).

7 Majority Harmonization Means Harm

One might argue that for the remaining 53.4% cases a majority resolution was appropriate. Just get rid of all singletons and adopt those annotations where the majority of the voters agrees (3 or 4 voters). Only the 2 voter cases would have to be dealt with on the basis of further discussions.

This presupposes that the majority perspective is the most valid one. We found out that, statistically, this is not the case. We reached that conclusion afterwards, i.e. after we have carried out harmonization on the basis of discussion (which was meant to clarify the reasons for disagreement in the first place).

Table 4 shows the frequency of adaptation decisions. For instance, the first row shows that the first annotator switched his opinion 22 times in the case when the three other annotators voted in a different way, i.e. voter 1 adopted his annotation decision. There are two variants of this: voter 1 canceled his annotation since the others have not approved it or voter 1 has not seen an annotation step the others have and now he adopts it.

annotator id	3 votes	2 votes	1 vote
1	22	16	10
2	12	4	11
3	8	9	8
4	12	13	10
	54	42	39

Table 4: Modifications per rater given counter raters

If majority vote proved to be superior over singleton votes, then, in general, the inclination to modify a decision should be dependent (increase) on the number of voters that stand in opposition to it. This means the more counter voters, the higher the probability of a modification toward that majority perspective. In order to test this, we specified as a null hypothesis that the modification decisions are independent (sic!) from the number of voters that stand in opposition to it: three voters change their mind quite as often as a singleton voter (i.e. with the same probability). Then we had independence.

If we can reject the null hypothesis, if the majority vote more often prevails, than the harmonization strategy *majority voting* has proved valid (and useful). If not, we have evidence that singleton opinions are quite as valid as majority decisions. We then have to discuss the status and consequences of such a finding, namely whether we should harmonize at all (see below).

We applied Fisher's exact test (in R) to the table 4 and get as a p-value 0.35 which obviously is not significant at any level (e.g. $p < 0.01$). The null hypothesis (independence, i.e. P(annotator's inclination of revision|number of counter raters) = P(annotator's inclination of revision)) cannot be rejected thus. This strengthens our claim that harmonization should not just be realized as majority voting.

8 Any Harmonization is Harm

We found that a single opinion might turn out to be as valid/strong as three opposing ones. But what does this mean for the harmonization idea? The majority harmonization strategy produces a gold standard that only seemingly represents the best choice - as we have seen. The discussion-based harmonization strategy on the other hand produces a gold standard where non-representative opinions are as frequent as representative ones (those of the majority of raters). Such a gold standard no longer represents the prototypical reader - an entity we would like to model.

As a consequence: we neither should harmonize by majority vote nor by discussion-based agreements. We should not harmonize at all (beyond the elimination of mistakes, of course).

No harmonization means that our systems could make use of all annotations in order to learn a distribution of opinions. We then could interpret the probabilities such a system would assign to a particular decision as an indicator of its prototypicality or prominence (visibility). It also would produce singletons which might represent interesting perspectives as well. This, of course, requires further investigation and further proof.

9 Related Work

There exists a couple of papers dealing with sentiment inference, see e.g. (Deng and Wiebe, 2015a), Rashkin et al. (2016), Klenner and Amsler (2016),Klenner et al. (2017). There are also some annotated resources, e.g. the MPQA corpus (Deng and Wiebe, 2015b), but all approaches that we know rely on a harmonization step. There are also a number of papers dealing with (mostly crowdsourcing related) annotation quality ((Plank et al., 2014), (Hovy et al., 2013), (Geva et al., 2019), (Sheng et al., 2008)). But none of these approaches argues in favor of a gold standard in the form of an decision distribution, as we do.

As a notable exception, Kenyon-Dean et al. (2018) present an interesting discussion on the disagreement of annotators for the sentiment labelling task. They demonstrate that bare averaging or simple heuristics, such as the majority vote,

should be avoided. In contrast to our work, they consider a task that includes only one label per utterance, whereas we focus on effects and relations, including multiple, possibly independent instances per sentence. Additionally, in their work, they investigate crowd-sourced annotations rather than observing the outcomes of the harmonization stage in the form of a discussion among the annotators as we do. However, we see many similarities in this paper to study sentiment annotation disagreement, but viewed from a different perspective. Also, we fully support the authors on their postulate to release corpora with annotations from all annotators, and withstanding from discarding data samples for which the agreement is low.

10 Conclusion

In this paper, we argued in favor of an annotation strategy that harmonizes as much as reasonable (in order to get rid of errors and annotation omissions), but otherwise leaves the distribution of annotator decisions intact. We provided first statistical evidence for such a strategy in the area of sentiment inference, where annotation decisions are far more subjective than, say, in PoS tagging. We are interested in models of a prototypical reader, we strive to model his/her understanding and we believe that training a system on the basis of a distribution of opinions better serves our purposes as if an artificially created gold standard was used.

Acknowledgements We would like to thank Noëmi Aepli, Sophia Conrad and Andreas Säuberli for their valuable contributions. This work is supported by the Swiss National Science Foundation under the project ID 105215_179302.

References

Lingjia Deng, Yoonjung Choi, and Janyce Wiebe. 2013. Benefactive/malefactive event and writer attitude annotation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–125, Sofia, Bulgaria. ACL.

Lingjia Deng and Janyce Wiebe. 2015a. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 179–189.

Lingjia Deng and Janyce Wiebe. 2015b. MPQA 3.0: An entity/event-level sentiment corpus. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, Denver, Colorado, May31–June5, 2015*, pages 1323–1328.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. ACL.

Manfred Klenner and Michael Amsler. 2016. Sentiframes: a resource for verb-centered German sentiment inference. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1–4, Paris, France. European Language Resources Association (ELRA).

Manfred Klenner, Don Tuggener, and Simon Clematide. 2017. Stance detection in Facebook posts of a German right-wing party. In *LSDSem 2017/LSD-Sem Linking Models of Lexical, Sentential and Discourse-level Semantics*.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. ACL.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany.

Victor S. Sheng, Foster J. Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD*, pages 614–622. ACM.